DOCUMENT RESUME

ED 049 008 RE 003 439

AUTHOR Murray, James R.
TITLE An Experimental Design For Summative Evaluation of Proprietary Reading Materials.
PUB DATE Feb 71
NOTE 57p.; Paper presented at the meeting of the American Educational Research Association, New York, N. Y., Feb. 4-7, 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS *Educational Research, *Evaluation Methods, Intermediate Grades, *Program Evaluation, Reading Instruction, *Reading Materials, Reading Research, Remedial Reading, *Research Methodology, Statistical Analysis

ABSTRACT

A summative evaluation design was developed as a framework for evaluating instructional materials in remedial reading. The paradigm includes the selection of (1) relevant variables for study and (2) the method of study. Two types of reading materials used in Chicago schools were studied--Cracking the Code (CTC) and the Mott Semi-Programmed Series in Language Skills (MLS). Random procedures were used to select the 36 classrooms studied (two classrooms at each of the fifth-, sixth-, and seventh-grade levels from schools in each of six school districts representing high, middle, and low socioeconomic levels). Teachers in these classrooms were randomly assigned to one of the programs for 1 month and asked to use the programs as supplements to regular instruction. Pretesting and post-testing results were compared. Among the conclusions were (1) that program effects are multiple, (2) that differences based on socioeconomic levels vary at different grade levels, and (3) that no simple decisions are possible regarding which of the programs is superior. Tables of analysis of variance results and references are included. (MS)

ED049008

U. S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

An Experimental Design

For Summative Evaluation of Proprietory Reading Materials

James R. Murray
Industrial Relations Center
The University of Chicago

RE003 439

Presented at the 1971 meeting of the
American Educational Research Association, New York

I. General Background

A serious and growing need is developing in American education for evaluation of educational programs, from the level of the specific textbook to the level of the general school system. The work of the Center for the Study of Evaluation of Instructional Programs at UCLA, the American Educational Research Association's sponsorship of monographs and symposia on problems in evaluation, and the extensive funding of local evaluation under the Elementary-Secondary Educational Act (Title III) each point up the increased professional awareness of the need for sound evaluation methodology in education. A number of papers have been written concerning the nature of the problems in evaluation and appropriate research methodologies. Traditional psychometric achievement testing, with its emphasis on individual differences, has been challenged as a paradigm or theory for measurement in program evaluation (Gagne, 1967, 1968; Cronbach, 1963; Tyler, 1967 and Stake, 1967). Scriven (1967) has raised a number of additional questions about the nature of educational evaluation and, for example, has challenged the appropriateness of using only the classical criteria of research designs in which explanation is the primary goal. This so-called "explanatory" research design is a strategy espoused in the well known paper of Cronbach (1963). It is in the context of such intellectual controversy that Westbury's (1970) finding, i.e. that actual curriculum evaluation research has not been reported in the educational literature, appears disappointing. It is hoped that the present research can contribute to the development of our ability to do evaluation.

The research presented here is intended to realize a meaningful and significant approach to the general problem of educational evaluation. This study was intended to have value beyond simply knowledge of the specific curricula studied here, Remedial Reading Programs. Many basic and pervasive problems are encountered in any educational evaluation research and how they are dealt with must have an effect on the quality of the program evaluation. The design used for this evaluation was intended to avoid some of those restrictions often encountered in educational research and thereby allow for more direct and meaningful applications.

The primary problem framework of this research can be stated rather simply. What information does a classroom teacher or school administrator need in order to decide which of several commercially available curriculum programs should be purchased and used? Currently, the information available is quite informal, e.g. the recommendation of teachers, the brochures or orientations given by sales representatives (usually rather devoid of facts) or simple common sense-experience which the teacher has acquired. Furthermore, individual teachers are rarely free to choose among all possible curriculum programs. Many states have state-adoption programs which limit schools to use only those materials which have been officially adopted. In some cases the school or school system has curriculum staff which serve a screening function for curriculum materials, or schools may arrive at group or administrative "policy" decisions about what materials will be acquired by the school itself, from which the individual teacher can then select. It is important to note that even when purchase decisions are school-based, the available informationfor decisions is still quite informal and generally intuitive.

The ideal situation from the perspective of the administrator or teacher appears to be one in which each publisher would make available extensive information on program outcomes as well as extensive cost information, e.g. materials cost, teacher training cost, usage time requirements, etc. This is not being done. Publishers contend that they have neither the financial wherewithal nor the technical skill to provide all of this information. What are the alternatives?

Although neither total cost nor outcome information is available, it is clear that the more pressing need is for facts about the outcomes or results of program usage. This type of information must be available for rational decision making in education and, joined with cost data, forms the only intelligent basis for efficient allocation of educational resources (Alkin, 1969). A suggestion that school districts themselves perform their own evaluation is not feasible. Wiley and Bock (1967) point out some of the relatively straightforward problems encountered in this approach, primarily arising from the limited experimental control possible in a single district. Some obvious constraints involving the financial limitation of school districts, parental resistance to pervasive and continual innovation in the schools, and teacher resistance add to the list of such difficulties.

A viable strategy for acquiring the necessary information seems to require the participation of independent investigators. University faculty or research institutes, supported primarily by numerous school districts in consort or independently funded, can provide the required technical competence, objectivity and capacity to utilize multiple school districts in exploring program outcomes. This evaluation program was undertaken

to "test", as it were, the viability of such a cooperative, inter-district model for curriculum evaluation.

Scriven's (1967) conceptual framework provides the terms "summative" and "pay-off" which can be used to describe the kind of evaluation needed for commercial programs. What is at the core of the decision problem from this perspective is the acquisition of knowledge concerning the outcomes or behavioral results due to the application of a curriculum program. This is a "blackbox" perspective in which performance or output is the focus rather than an attempt to provide a detailed explanatory specification of instructional process as is apparently proposed by Cronbach (1963) or Bormuth (1969). However, it is not enough even to take sides on that issue. What is also needed in order to do evaluation research is a working framework or paradigm which points at 1) relevant variables for study and 2) the method of study. The following considerations were used as such a framework, and provided a basis upon which the present investigation was designed.

The Working Framework

1) The Variables:

   1. The Educational Program (curriculum) analysis:
      a. What is the content?
      b. How is it used?
      c. Who is to use it?
      d. Who is to receive it?
   2. The measurement of outcome:
      a. What are the skills or knowledge directly taught in the program?
      b. What are the general skills or knowledge built upon the direct skills?

3. What are the properties of schools which may be related to program outcome or effects?
4. Are there properties of students which may be related to program outcome or effects?
5. What range or extent of applicability of information is needed or desired?

2) The Method:

The most appropriate and direct method available for obtaining information on the comparative effects of educational programs is the experimental method. It allows the investigator to actively manipulate and control different variables of interest. The theory of experimental design, as developed by R. A. Fisher, is built specifically on a procedure called randomization. This procedure guarantees the validity of inferences about the effect of influences of experimental treatments. It should be clear that this property of inferences is very badly needed in education evaluation. The experimental paradigm also,or perhaps primarily, has furnished an extensive basis for analyzing resultant data and making inferences based on such data. There has been an excellent critique of the problems in the use of experimental design in education (Campbell and Stanley, 1963). Wiley and Bock (1967) show some aspects of at least one way randomization can be appropriately done, i.e. on the level of the classroom. Hopefully, demonstration of the application of experimental strategies to evaluation can facilitate the practice and development of curriculum evaluation.

The primary importance of the questions listed under "Variables" in the Working Framework is that answers to them can specify the relevant

aspects of an educational program in terms of the parameters which could effect program results. Given these properties of a program, an experimental design could be constructed to yield important information for use by the prospective decision maker. These considerations were specifically applied to the Remedial Reading programs evaluated in this study and will be reviewed below.

Measurement of Program Effects

There are two properties of any measurement procedure used in evaluation which are critical. Both of these properties basically involve "validity" considerations, as opposed to the usual concern with the statistical "reliableness" of measures. First, there must be an acceptable correspondence between the instructional content of the program(s) and the behavior or performance observed in the measurement procedure. The degree of this correspondence cannot be itself measured absolutely, but it can be judged qualitatively. Gagne's (1969) term distinctiveness may well apply here. The second property could be referred to as completeness. What is of concern here is the scope or breadth of observations of phenomena which are "indirectly" related to the immediate content of the program(s). For example, a measurement procedure which included "thought" questions based on an arithmetic program would be more complete than one which only included simple computational exercises. A classical learning paradigm would refer to these more"complete" observations as measures of transfer or of response generalization.

The area of instruction investigated here is that of reading. The remedial nature of these curriculum materials imposes a very significant additional constraint on the content of the programs. The emphasis on the so-called "decoding" process, i.e., generating phonetic representation of the written text, is evident in both of the programs studied here. The commonality and inclusion of such letter-to-sound training is due to the belief that:

1. This skill is clearly a prerequisite for the real business of reading, comprehension of meaning;
2. Many children cannot read and comprehend meaning in written materials because they are not able to decode from letter to sound; and
3. Therefore, they must be trained in that skill.

The reasonableness of the first and second parts of the above rationale is not completely known (Desberg and Berdiansky, 1968, Levin and Gibson, 1968). It could be argued that it is irrelevant to the evaluation measurement problem. What must be done, nevertheless, is measurement of these instructional skills because they are taught by the curriculum and therefore relevant to evaluation.

Two measurement devices were used which are related to the decoding, or word attack, skills. The Letter-Sound Correspondence Test (LSC)[1] and the Bond-Clymer-Hoyt Silent Reading Diagnostic Tests[2] (SRD) were chosen, not only for their distinctiveness, but for the fact that they are group administered tests, a very necessary attribute. The LSC test is based on linquistic studies of English orthography and the basis for the measurement procedure is given in Venezky, et al (1969). The SRD test can perhaps be described best by a list of the subtests used:

1. Syllabication,
2. Root Word Location,

---

[1]Under development by R. Venezky, R. Chapman, and R. Calfee of the Research and Development Center for Cognitive Learning at the University of Wisconsin.

[2]Published by Lyons and Carnahan, Chicago.

3. Word Elements (Sound to Letter),
4. Beginning Sounds (Sound to Letter),
5. Rhyming Sounds (Sound to Letter),
6. Letter Sounds (Sound to Letter).

The second property of evaluation measurement, completeness, was realized here through the use of the Iowa Silent Reading Tests[1] (ISR). The ISR test is primarily a test of comprehension skills, although an analysis of the sources of information used in the test shows the test to be quite complex (Bormuth, 1968). The test is group administered and a traditional, widely used test of reading. There are two primary reasons for including such a test in the evaluation measurement. First, effective comprehension of written material is the basic goal or target of reading instruction; it is the final behavioral objective. Therefore, in a fundamental sense, no instructional program for reading, whether remedial or not, can be evaluated without some measurement of comprehension behavior. Second, the assumption common to both instructional programs, i.e., the key role of decoding/word attack skills in remedial reading instructions, forces one to go beyond the measurement of only letter-sound knowledge. The situation which must be avoided is one in which instructional effects are investigated on letter-sound knowledge, but there is no evidence collected regarding instructional effects on comprehension skill <u>through improvement in letter-sound knowledge</u>. The assumption made in the materials in order to arrive at a remedial program must not remain an assumption in evaluation, but become an hypothesis subject to empirical examination. That is, does improvement in letter-sound-correspondence knowledge lead to increases in comprehension skill?

[1]Published by Harcourt, Brace and World, Inc.

Instructional Materials

The commercial materials examined in this research are the Mott Semi-Programmed Series in Language Skills (MLS), published by the Allied Education Council, and Cracking the Code (CTC), published by Science Research Associates. A step which must be taken is an analysis of the content, methods and goals of the two programs.

The Cracking the Code materials are designed to teach children basic letter-sound patterns using a deductive approach. This method, often called, "linguistic word attack," presents regular grapheme-phoneme correspondences in several words. It is hoped that, by practicing such words, the child will either:

1. Discover the letter-sound rules and then use them inductively, or
2. Figure out new words by analogy, using known patterns.

The core of the program is the workbook and is divided into twelve sections, each with a corresponding section in the accompanying reader. The reader is designed to provide practice using the word patterns that have been learned from the workbook. These patterns are introduced according to their "frequency of occurrence in writing" and "ease of discovery." Infrequent or difficult patterns are introduced near the end of the program. However, many of the word patterns introduced in the same lesson can be easily confused. For example, lesson 12 introduces the patterns ight, igh, eigh, ought, and aught.

Since reading is defined as a process of decoding writing into sound, word recognition skills are taught and vocabulary and comprehension skills

are ignored. But even within this restricted framework, problems outside of grapheme-phoneme correspondences have been handled superficially, where they have been handled at all. For example, the word recognition skills of syllabification and morphological division receive very little systematic attention.

The book of readings accompanying the workbook presents words containing the sounds introduced in the workbook. The selections begin with quite easy words presented in an extremely "linguistic" format ("it was odd to run into a bug in the lap of a Dupenpox on top of a hill") (p. 10). However, they quickly progress into more conventional readings. There is no noticeable change in the difficulty of the vocabulary or syntax of the stories from page 30 to the end of the book (page 215). However, this is an impressionistic analysis; readability formulas have not been applied.

The Mott Semi-Programmed Series in Language Skills materials are more eclectic in both content and approach. They include exercises in writing as well as in all phases of reading. Comprehension, vocabulary and word recognition skills are taught. The program includes many practical applications of reading such as reading labels and newspapers. The MLS materials may be divided into the first six and the last four books. The first six books teach letter-sound correspondences. They are roughly comparable to the CTC series. For the most part, an inductive phonics approach is used. The last four books present extensive reading and vocabulary and more advanced exercises on word recognition skills such as syllabification and morphology.

The MLS materials combine inductive and deductive methods for teaching grapheme-phoneme correspondences. Typically, a word is introduced which contains the pattern to be taught. For example, if the "ase" pattern is to be taught, the word "case" is used. By changing the first letters, several words are formed with this pattern, ("lace", "race", "place"). Such an approach requires the ability to blend letter sounds into words.

The sequence of the first six books is roughly comparable to CTC. However, much additional material such as stories and word studies are added in books five and six. The pacing is indeterminate since each child supposedly proceeds at his own speed, although MLS seems to be slower than that of CTC. CTC may present several deductive patterns simultaneously, but MLS will present only patterns. The MLS program tends to present sounds in units. For example, the hard and soft sound of "c" and "g" are presented in sequence, the three sounds of "oo" are in sequence, the three sounds of "es" are presented in sequence. "Eu" and "ew" representing the same sounds are presented in sequence. There is some review provided; it appears in large but irregular intervals.

Books seven, eight, nine, and ten of MLS present many lessons in reading. Several, listed under "American Scene" have very practical applications such as reading labels, newspapers, magazines, etc. There is also extensive vocabulary study in "word study." In addition, the following topics, which may be considered word recognition skills are treated in detail: book seven--compound words, prefixes and suffixes, syllabification; and book eight--synonyms, antonyms, and homonyms.

This review of the content and method of the programs provides a basis for determining an appropriate domain for evaluation. This review

must also be placed in the context of the avowed goals and limitations of the programs as presented by the publishers.

1. Both programs are intended for use with "non-pathological" remedial readers. Only CTC is more restrictive with its focus on readers with only "decoding" difficulties.
2. Both programs are intended for use with children in the middle grades, i.e., five through nine.
3. Both programs are intended to be used by the classroom teachers.
4. Both programs lack detailed placement or diagnostic procedure for use with the programs.
5. Both programs are introduced to teachers primarily through an accompanying teacher's manual. Orientations given by sales representatives are 30 to 90 minutes long and focus on explaining the manual.
6. Both programs are designed to be supplementary, in that they are not intended as the sole material to be used for reading or language skill instruction.

Instruction and the Schools

The basic question which must be answered here is, are there any properties of schools which can mediate the influence of the instructional program? Certainly there is a non-trivial problem in specifying which properties are truly associated with schools as units versus simply aggregate qualities of pupils in the schools. Correlations between average student I.Q., say, and average teacher salary or education level, need not imply a reductibility of one to the other. Important characteristics of neighborhoods, which give rise to both average student I.Q. level and to teacher

salary level, can be sources of influence common to these conceptually independent phenomena and therefore lead to non-zero correlation between them.

Scriven (1967) makes the point that whenever a set of materials or an instructional program is used in the classroom, the program itself is not only realized, but is incorporated into the entire instructional sequence which a teacher implements. Thus, the instructional program for the students consists of the materials in the hands of the teacher. Furthermore, the general instructional activity of teachers is a part of the educational practices of the teacher's school or school system. Therefore, one would expect instructional practices of teachers to differ in association with relevant differences among schools. Finally, the single, most pervasive property which can be associated with schools is its socioeconomic status as a unit. Primarily financial,but also concomitant educational and occupational, attributes of the neighborhoods in which schools operate determine and constrain in various ways the educational practices of the local school.

The fact that the MLS materials were originally developed for use in a midwest industrial town which is noted for its poverty and illiteracy, leads us to anticipate that this program may have a greater effectiveness in poorer rather than wealthier schools. Conversely, the CTC materials are derived from materials which have had a good deal of success in suburban school systems. It appears to be a reasonable question as to whether CTC will be as effective as MLS in poorer schools. For both of these programs, the possibility of differential effectiveness is based on considerations of the practices and resources of the schools themselves. In poorer schools, it is not simply that they may have a larger number of

deficient readers,but it is that they have almost no resources, either in staff or equipment, to deal with these remedial children. A single visit to a wealthy school system can demonstrate the extensiveness of resources available there for special problem students. These differences among schools make up significant behavioral systems in which new materials are utilized.

Student Characteristics and Instruction

The socioeconomic properties of schools, it has been mentioned, are associated with the aggregate properties of students. The major investigations of socioeconomic status and educational variables have considered the individual pupil as the unit of study. Jensen (1969) stated, "The relationship between SES and IQ constitutes one of the most substantial and least disputed facts in psychology and education." Furthermore, Whiteman and Deutsch (1968) found substantial correlations between socioeconomic status and reading performance. Their findings include the well known substantial correlation between reading performance and IQ, and therefore, the concomitant joint association of these two variables with SES. These results are all based on individual pupil characteristics.

Although it is conceptually problematic, it is fortunate on the practical level that controls for SES properties of schools implicitly control for SES properties of pupils. The conceptual problem centers around the determination of which agent, school vs. pupil,is the basic or primary vehicle for the influence of SES on program effectiveness. However, again on the practical level, this conceptual problem may in fact not be a relevant problem. American society is such that pupil and school, via at least a common neighborhood, have highly similar SES qualities. The significant

implication is that any instructional program, for use by classroom teachers, will invariably be inserted into a classroom situation in which these SES properties are jointly in effect. Thus, evaluation of program effectiveness can yield sufficient information by simply treating school-plus-student as a functional unit, i.e. ignoring the issue about which is more "important."

There is an additional variable of students which is relevent here. The variable of age, or more directly,grade of the student is important because the materials are intended for use with children who are beyond grade four. Such a wide domain of use forces one to question the uniformity of program effectiveness over grade levels. In the first place, deficient readers in the higher grades (above grade 6) have not only failed more but may have developed quite different strategies for dealing with their problem than their younger counterparts. Also, the effects of repeated failure on attitudes and motivations of older students certainly cannot be ignored in remedial instruction. Thirdly, the cognitive structures which students bring to bear in new learning experiences certainly should be expected to differ by grade level. Gagne (1968) has outlined alternative ways in which these differences can arise and effect instructional success, and Cronbach and Snow (1969) have described a phenomenon which is related to this issue, the Aptitude-by-Treatment Interaction (ATI).

The Population: Range of Application

It must be quite explicitly realized that the essential goal of commercial materials evaluation is to investigate the effectiveness of programs as they are normally to be used. There are two primary attributes of an evaluation study in this regard. First, the "treatment" or program

administration which is realized in the study must be directly related, i.e., highly similar, to the programs as they will be administered in the normal, non-research setting. Second, the results of the research must be applicable to as wide a range of potential consumers as possible.

These two facets of the inferential goals of evaluation were accomplished in this study by:

1. Simulating in the study, as thoroughly as possible, the normal process of materials introduction and usage as obtains in the commercial setting; and
2. Specifying a population of schools from which a true random sample could be drawn for participation in the study.

Both of these procedures are described in the procedure section of this document. The point to be made here is that without both of these procedures the results or inferences of an evaluation study will be of limited value because:

1. The nature and conditions of program administration will not be the same, or highly similar, between research and actual usage;
2. The kinds of school/pupil milieus or situations in which the programs have certain effects will not be practically specifiable and generalizable to potential consumers.

Summarizing the above considerations for the evaluation of the MLS and CTC remedial reading programs, the following decisions were made about the research design.

1. Randomization, i.e., true experimentation, would be used for program assignment to classrooms.

2. The program materials would be studied (used) in actual classroom situations, accompanied by the same procedures used by the publishers with normal consumers.
3. The socioeconomic status of schools, and thereby pupils, would be studied in relation to program effectiveness.
4. The grade level of students using the materials would be studied in relation to program effectiveness.
5. True random sampling of schools from a specified population (sampling frame) would be done.

II. Procedure

Design of the Study

The two remedial reading programs evaluated in this study were the Mott Semi-Programmed Series in Language Skills (MLS), and Cracking the Code (CTC). Both programs utilize the linguistic approach to reading instruction and are intended for use in the fourth through sixth grades and up. The two programs differ in mode of presentation. The MLS employs a programmed instruction format for word-attack skills and comprehension. The CTC, on the other hand, relies solely on teacher-guided word-attack (decoding) exercises and utilizes prose-reading solely for practice. Neither the MLS nor the CTC are claimed to be innovations in the teaching of reading. Both programs involve principles (e.g., linguistic approach and programmed format) present in other currently available reading programs. However, little research substantiating the effectiveness of these principles has thus far appeared in the literature.

The design of this study has two distinct parts. The first involves the selection of the schools and classrooms for participation in the study. The second involves the assignment of treatments or materials to the classroom.

The classrooms actually used in this study were obtained by a process of sampling known as stratified random sampling. From a listof 250 communities and Chicago neighborhoods published by the Chicago Association of Commerce and Industry, the major incorporated areas (and neighborhoods within Chicago) in the Standard Metropolitan Statistical Area of Chicago

were divided into three groups based on the median family income, average home value, and assessed property valuation of each area. The three groups were, for our purposes, labeled or defined as socioeconomic levels high, middle, and low. Separately within each of these groups of 83 areas or neighborhoods, 18 areas were randomly selected for contact. The goal was to obtain six areas at each SES level for inclusion in the study. Fortunately, each area was served by a single school district, and it was these concomitant school districts that were contacted for participation.

The second part of this study design involved randomly assigning the treatment conditions to classrooms within each district. It was generally the case that most schools have only two classrooms at each of the middle grade levels, i.e., fifth, sixth and seventh grades. Because of our desire to use classrooms from the same school, and in general having only two classes at each grade level, the design chosen involved assigning only two of our three materials conditions (this includes a control) within each grade within each school district. Since we wanted to study the effects of both grade and SES on treatment effectiveness, we adopted a plan for randomly assigning two treatment conditions which balanced the influence of grade, SES, and treatment over each other. The design is best represented by Table 1, and is a partially balanced incomplete block (PBIB) design (Kempthorne, 1952). There are over 2500 Ss included in this study.

Testing

The classrooms chosen for study were administered our reading test battery in the classroom as a group. The tests were all group tests and designed for administration by non-specialists in either the field of reading or psychological testing. The battery generally required three hours of classroom time for administration, with a break given to the students about halfway through the battery. All of the pretests were administered by staff members at the Industrial Relations Center. The posttesting was done primarily by Industrial Relations Staff, but approximately one-fourth of the classrooms were tested by the classroom teacher. Care was taken to spread the teacher-tested classrooms over SES levels and treatments.

Three measurement instruments were used in this study:

1. The Iowa Silent Reading Tests--Form CM
2. The Silent Reading Diagnostic Tests--Recognition Techniques
3. The Letter-Sound Correspondence Tests--Version II

Materials Presentation

The teachers who were randomly assigned to use either of the remedial materials were given an orientation to their respective materials during the period of pretesting. Included is an outline followed by the orientors in the general portion of the introduction to the research which all teachers were given. (See Appendix A)

TABLE 1

The Randomized Incomplete Block Experimental Design

	Grade 5		Grade 6		Grade 7	
	Classroom 1	Classroom 2	Classroom 1	Classroom 2	Classroom 1	Classroom 2
SES I District						
1	Mott	SRA	Mott	Control	SRA	Control
2	Mott	SRA	SRA	Control	Mott	Control
3	Mott	Control	Mott	SRA	SRA	Control
4	Mott	Control	SRA	Control	Mott	SRA
5	SRA	Control	Mott	SRA	Mott	Control
6	SRA	Control	Mott	Control	Mott	SRA
SES II District						
7	Mott	SRA	Mott	Control	SRA	Control
8	Mott	SRA	SRA	Control	Mott	Control
9	Mott	Control	Mott	SRA	SRA	Control
10	Mott	Control	SRA	Control	Mott	SRA
11	SRA	Control	Mott	SRA	Mott	Control
12	SRA	Control	Mott	Control	Mott	SRA
SES III District						
13	Mott	SRA	Mott	Control	SRA	Control
14	Mott	SRA	SRA	Control	Mott	Control
15	Mott	Control	Mott	SRA	SRA	Control
16	Mott	Control	SRA	Control	Mott	SRA
17	SRA	Control	Mott	SRA	Mott	Control
18	SRA	Control	Mott	Control	Mott	SRA

The main emphasis in the materials orientation given to the teachers was a description of the materials, how a teacher was to use the materials and, most important, a review of the teachers' manual and how it was to be used. It was not one of the goals or practices to present to the teachers a theory or new concept of teaching reading to poor readers. Our main goal was to get the teachers into the manuals and help them with questions. It was expected, or hoped, that the manuals would carry the primary burden of teacher instructions. We stressed to the teachers that they were to contact us if they wanted assistance and also that we would follow-up with them in January. Finally, since no placement or diagnostic procedures accompanied the materials, the teachers were instructed to use the materials with any student they decided, by whatever means, might benefit from the instruction.

III. Results

The nature of the assignment of the reading materials to students in this study was based on the classroom as an administrative teaching unit. The grouping of students into classes for instruction will usually be reflected in similar performance among children in the same class. This similarity of performance of students, as grouped by classrooms, must be directly accounted for in the analysis of the effects of the programs being studied. Thus, instead of there being 2,500 observations for analysis of program effects in this study, i.e. the number of pupils measured, there are only 124 observations, i.e. the number of different classrooms actually measured.

Wiley and Bock (1967) provide relevant data as well as a rationale for treating the classroom as a unit of analysis, and the reader is referred to that paper for a more thorough elaboration of the strategy. The primary goal of the analysis reported here is to assess the performance effects of the two reading programs. There are four general aspects of the results presented here:

1. Description of the measures;
2. Distribution of program usage;
3. Analysis of program effects by independent variables, e.g. main effects and interactions
4. Analysis of program effects by dependent variables, e.g. over skills.

1. Description of the Measures

The standard deviations and reliabilities for the following six scores are based on the pooled within-classroom variability, e.g. the student's score minus the average for his classroom. These are presented in Table 2.

2. Distribution of Program Usage

In the procedure section, it was pointed out that the classroom teachers were assigned one of the two reading programs. They were free to determine the extent of use of the materials in their own classroom. This teacher option resulted in the frequencies of actual student participation in the programswhich are presented in Table 3.

These frequencies show two phenomenon. First, there is a greater usage of materials in the lower economic group than in the higher, a not too surprising result. Second, there is a greater use of the Mott (MLS) materials than the SRA materials within similar classroom categories. This may have arisen from the apparent differential participation of the teacher in using the materials, with the MLS being semi-programmed.

3. Program Effects-Overall Multivariate Comparisons

The preceding data on the differential usage of the program materials does not in itself complicate analysis of performance differences. However, the fact that for both groups of program classrooms there were some students within individual classrooms who did not use the materials, while some students in the same classroom did use them, is somewhat problematic. The use of the classroom as a unit of analysis for comparing program effects usually rests on the fact that all of the students in the classroom are treated similarly with respect to the instructional

TABLE 2

STANDARD DEVIATIONS AND RELIABILITY COEFFICIENTS

OF THE RESPONSE MEASURES

	Pretest		Posttest		Number of Items
	S.D.	r_{xx}	S.D.	r_{xx}	
Letter-to-Sound Test	8.55	.872	7.78	.858	50
Syllabication & Root Word Tests	6.52	.842	5.88	.831	54
Sound-to-Letter Tests	11.50	.864	10.53	.884	120
Paragraph Comprehension Tests	10.01	.873	11.37	.900	90
Vocabulary Tests	6.49	.759	7.11	.805	54
Sentence Meaning Test	3.51	.532	3.96	.710	27

TABLE 3

NUMBERS OF PUPILS WHO RECEIVED THE MATERIALS

FOR EACH GRADE IN EACH

SOCIOECONOMIC LEVEL

SES	GRADE	Mott	SRA
1	5	41	9
	6	30	19
	7	14	5
	Sum	85	33
2	5	74	33
	6	64	53
	7	48	54
	Sum	186	140
3	5	100	20
	6	59	50
	7	63	62
	Sum	222	132

programs being studied. This is not the case here. An analysis which did, nevertheless, average the scores of all students within a classroom, and thereby ignored actual usage patterns, could obscure the detection of actual program effects.

The analytic procedure which was chosen here attempted to incorporate both the classroom as the basic unit and the fact that there are within classroom treatment differences. The technique used to do this involved doubling the number of measurements for each classroom. The two sets of measures associated with each classroom consisted of the pre- and posttest averages for, first, the group of students who did not receive the instructional materials and, second, the group of students who did receive the materials. This results in a 24 element response vector for each classroom, made up of the six reading scores for both pre- and posttests each for both treatment sub-groups within the classroom. This allows for the fact that the performance of the two groups of students are correlated as a result of their being in the same classroom. Tests of significance in an analysis of variance will thereby not be invalidated because the error covariance matrix can reflect the intraclass correlation among the subgroup scores.

There are two facets of the program effects which can be readily examined using this arrangement of the data. The first involves comparing the performance of only the students who received materials across classroom factors, e.g. grade or SES by program type interactions. The second set of comparisons involves examining the within classroom differences

between the treatment subgroups and studying the relative differences over classroom factors.

The multivariate F-ratios for the ANOVA corresponding to various treatment effects are presented in Tables 4 and 5. The incomplete and partially balanced nature of the research design renders the effects correlated. The order in which the F-ratios are performed, since the size of the mean squares are effected, is important. Also, there are a different error terms for various sources of variance. The following structure was used for the ANOVA here. All terms are based on the elimination of preceding sources of variance.

Source	df
Grand Mean	1
(A) SES	2
School (error for A)	18
(B) Grade	2
School x Grade (Lin) (error for B)	18
(C) Treatment	2
School x Treatment (error for C)	18
(D) SES x Treatment (MLS-SRA)	2
Grade x Treatment (MLS-SRA)	2
SES x Grade x Treatment (LMS-SRA)	4
Residual (error for D)	43

Table 4 shows the F-ratios for the vector contrasts of posttest measures, corrected for pretests, for the MLS versus SRA program comparisons. Single degree of freedom comparisons are presented rather than pooled tests. The last F-ratio, SES (Quad) x Grade (Quad) x Program, is found to be

significant. This would imply that the comparative effects of the programs depends on the Grade and SES levels of the classrooms in which the programs are used. This will be examined in more detail below.

Table 5 shows the F-ratios for the vector contrasts based on the within classroom subgroup differences in posttest performance, adjusted for the pretest differences. This table indicates that only the overall program differences are of importance to performance.

TABLE 4

MULTIVARIATE F-RATIOS FOR COMPARISONS OF

PROGRAM PARTICIPANTS USING PRETESTS AS COVARIATES

Source[1]	F	P[2]
Program	3.534	**
SES (lin) x Program	0.691	NS
SES (Quad) x Program	0.521	NS
Grade (Lin) x Program	0.384	NS
Grade (Quad) x Program	0.746	NS
SES (Lin) x Grade (Lin) x Program	1.649	NS
SES (Lin) x Grade (Quad) x Program	1.505	NS
SES (Quad) x Grade (Lin) x Program	0.531	NS
SES (Quad) x Grade (Quad) x Program	2.709	**

(1) df = 6,43 for each F-ratio

(2) * = P < .05

** = P < .01

*** = P < .001

NS = Not significant

TABLE 5

MULTIVARIATE F-RATIOS FOR COMPARISONS OF WITHIN CLASSROOM PROGRAM VS. NO PROGRAM SUBGROUPS ON POSTTEST MEASURES, PRETESTS AS COVARIATES

Source	F-Ratio	P
Program	3.753	***
SES (Lin) x Program	0.394	NS
SES (Quad) x Program	1.300	NS
Grade (Lin) x Program	1.120	NS
Grade (Quad) x Program	1.187	NS
SES (Lin) x Grade (Lin) x Program	1.021	NS
SES (Lin) x Grade (Quad) x Program	1.184	NS
SES (Quad) x Grade (Lin) x Program	1.031	NS
SES (Quad) x Grade (Quad) x Program	1.312	NS

TABLE 6

RESULTS OF UNIVARIATE F-RATIO TESTS FOR COMPARISONS OF PROGRAM PARTICIPANTS USING PRETESTS AS COVARIATES

Source (1) \ Measures	Sound to Letter	Letter to Sound	Vocabulary	Root Word	Sentence	Paragraph Meaning
Program (Pgm)	NS	*	NS	*	*	NS
SES (Lin) x Pgm	NS	NS	NS	NS	NS	NS
SES (Quad) x Pgm	NS	NS	NS	NS	NS	NS
Grade (Lin) x Pgm	NS	NS	NS	NS	NS	NS
Grade (Quad) x Pgm	NS	NS	NS	NS	NS	NS
SES (Lin) x Grade (Lin) x Pgm	NS	NS	NS	NS	*	NS
SES (Lin) x Grade (Quad) x Pgm	*	*	NS	*	NS	NS
SES (Quad) x Grade (Lin) x Pgm	NS	NS	NS	NS	NS	NS
SES (Quad) x Grade (Quad) x Pgm	**	NS	NS	NS	*	NS

TABLE 7

OF UNIVARIATE F-RATIO TESTS FOR COMPARISONS WITHIN

CLASSROOM PROGRAM VS. NO PROGRAM SUBGROUPS POSTTESTS

Source	Measures: Sound to Letter	Letter to Sound	Vocabulary	Root Word	Sentence	Paragraph Meaning
Program (Pgm)	*	*	NS	NS	*	NS
SES (Lin) x Pgm	NS	NS	NS	NS	NS	NS
SES (Quad) x Pgm	NS	NS	NS	NS	NS	**
Grade (Lin) x Pgm	NS	NS	*	NS	NS	NS
Grade (Quad) x Pgm	NS	NS	NS	NS	NS	*
SES (Lin) x Grade (Lin) x Pgm	*	NS	NS	*	NS	NS
SES (Lin) x Grade (Quad) x Pgm	*	NS	NS	*	*	NS
SES (Quad) x Grade (Lin) x Pgm	NS	NS	NS	NS	NS	*
SES (Quad) x Grade (Quad) x Pgm	NS	NS	NS	NS	NS	NS

4. Program Effects-By Skill Measures

The multivariate tests presented above provide an assessment of the effects of certain factors on the entire vector of measures. This can be taken as the initial evidence of noteworthy differences which can be further examined by means of the univariate F tests, which are presented in Tables 6 and 7. These tables summarize the analysis by simply indicating, for each source of variance and for each measure, the outcome of the statistical tests.

Tables 6 and 7 present the corresponding univariate results of the two multivariate analyses presented above. These results are presented simply in terms of whether or not a variable has a significant F-ratio, ignoring the overall multivariate F-tests. It is clear that there are some single variable tests which are significant even though the multivariate tests are not significant.

Figures 1 through 12, in Appendix B, present the plots of the average classroom performance on these two sets of measures. These plots are of the posttests, corrected for pretests, i.e. the residuals which the analysis of covariance F tests are based on. This information is useful in examining the nature of the univariate tests presented in Tables 6 and 7.

IV. Discussion

The primary purpose of this research has been to demonstrate the application of an approach to summative evaluation. The development of

the evaluation design involved choosing certain variables, e.g. grade level and socio-economic level, which might effect the performance of classrooms in which the curriculum materials were used. The intention of this design is to provide information which school personnel can use in deciding which of many possible materials might yield maximum learning in their school system. The results of the data analysis presented here do indicate certain inferences about program effects. These results, because of the study design, should extend, on the average, to schools in and around Chicago and to the teaching circumstances which prevail in such schools.

One aspect of the results which are relevant to choosing between the programs used in this evaluation are those comparing the program groups only (Tables 4 and 6). The high order multivariate interaction effect implies that decisions should be made based on the economic characteristics of the school as well as consideration of the grade in which the program might be used. Furthermore, the univariate F-tests, corresponding to this source of variation, suggest that the performance variables of sound-to-letter knowledge and sentence meaning knowledge are effected most. Figures 1 and 6 indicate the complexity of the relationship among these factors. For example, in SES II, the SRA program is superior in grade 5 but not in grade 7, while the opposite relation holds in SES III. Because of such significant reversals in performance, simple decisions about "which program is superior" seem infeasible.

The use of multiple measures of performance focuses attention on the variable nature of program effects. Several of the univariate F-ratios are significant for interaction components, as presented in Table 6. This appears to be a quite instructive result in terms of the decision

practices often followed. Programs do not have single effects but multiple ones. The issue for the decision maker therefore entails choosing the target variables which require curriculum materials, e.g. the cognitive goals of instruction. Given these, phrased as analytically precise as possible, then data such as presented here are useful. The simple model, presented on page 5, of the hierarchy of reading skills could be used to develop priorities among choices based on the statistical results. For example, rather than considering the programs to be essentially similar because of the lack of significant differences on the comprehension measure, emphasis can appropriately be given to the sound-to-letter skill effects associated with the programs.

Such diversity of effects over variables can also be seen in Table 7. The information to be gained by examination of these program vs. no program differences within classrooms concerns the relative advantage of having a remedial program at all. For these measures, a negative sign indicates the program subgroup is superior in performance, after pretest differences have been adjusted for, relative to the no-program subgroup. For example, the sound-to-letter skill measure shows the MLS to have a greater net effect than SRA in SES I grade 5 but not grade 7, which is reversed in SES III. Again, this is the kind of skill which appears important in remedial reading instruction.

Although there are indications of SES related differences in program effects, it is somewhat surprising that there are not more striking differences. An explicit goal of this design and analysis was to randomly sample schools based on known economic differences and to restrict the level of analysis to the classroom as the basic unit. It may be that

studies of SES effects on behavioral phenomenon have been somewhat misleading as to the size of such effects. If schools are chosen for their extreme poverty and racial composition and the within classroom rather than between classroom variability is used as the error variance estimator, quite different results will obviously arise. Careful attention to design issues appears necessary for objective and rigorous evaluation to be achieved.

Appendix A

RESEARCH ORIENTATION

Hello, my name is _____ and I'm on the Research Staff at the University of Chicago and this is _____ also of the University of Chicago.

Today, we will be telling you about our research study of remedial reading materials and also how to use these materials.

Your principal(s) and superintendent have shown a desire to have your classes participate in our study. We are testing your children now, as you know.

After I tell you about the design of the study, we will talk about how the different materials are used. Before I begin telling you about the study, are there any questions?

We will be together for a little over two hours.

TELL ABOUT STUDY:

1. Community list
2. Random selection by SES level
3. Comparison of Mott <u>with</u> SRA <u>with</u> Control
4. Also, look at developmental trend or unequal successfulness of material depending on age
5. <u>SES</u> by <u>grade</u> by <u>materials</u>
6. Stress complete random assignment (selection of school <u>and</u> classrooms and assignment to treatment)
7. Post test in May

THINGS THAT APPLY TO ALL TEACHERS:

We are not evaluating you, we are evaluating the materials and your classroom will be one of thirty-six using a particular method.

The two sets of material are supplementary, i.e., they do not replace your basal material.

Who in your class will use this material is up to you:

You should decide who in whatever way you would normally decide to whom you will give extra help or special work.

The tests we are giving are only for research purposes and not diagnosis or assessment of who is remedial.

When you will start any student in this material is up to you:

Do this in whatever way works best for you and the students.

How much any student will use the material is up to you:

You can put some students entirely into this material until they are through it, or give it to them in addition to the basal work, or alternate between the two.

Key point is that business as usual, simply use these materials with problem students who would need special work anyway.

We are beginning each classroom with five sets of student materials. When you see that you will need more, simply notify us how much more you will need and we will send it to you. Do this by calling or writing to:

Mrs. _____ - 753-2025

Evaluation Research Division

Industrial Relations Center

1225 E. 60th Street

Chicago, Illinois 60637

Tell her your name, school and its address and how much of what material you will need.

We will contact all of the teachers using materials during December in a follow-up on any difficulties that you may have had using the material.

The only requirement, as far as we are concerned in using the materials is that <u>if</u> you are going to give <u>any</u> student <u>any</u> special <u>extra</u> (beyond the basal) help, we ask you to first use these materials with him. That is, no matter what else you do with your slow readers, we ask that you start them in these remedial materials (to whatever extent you feel is appropriate). If these materials do not work with some students <u>after real effort</u>, then, of course, discontinue their use of them.

In terms of the tests we are administering, we do <u>not</u> want to tell you <u>what</u> tests they are until <u>after</u> our post-test. The test results will be fed back to the school, though. We do not want to add anything out of the ordinary to how students are handled here.

Also, please do <u>not</u> actively try to find out what each other is doing (across methods--not within methods). This may contaminate the purity of the comparison.

Appendix B

Figure 1.

Program Posttest Means Adjusted for Pretests

for SES x Grade x Treatment: Letter-to-Sound

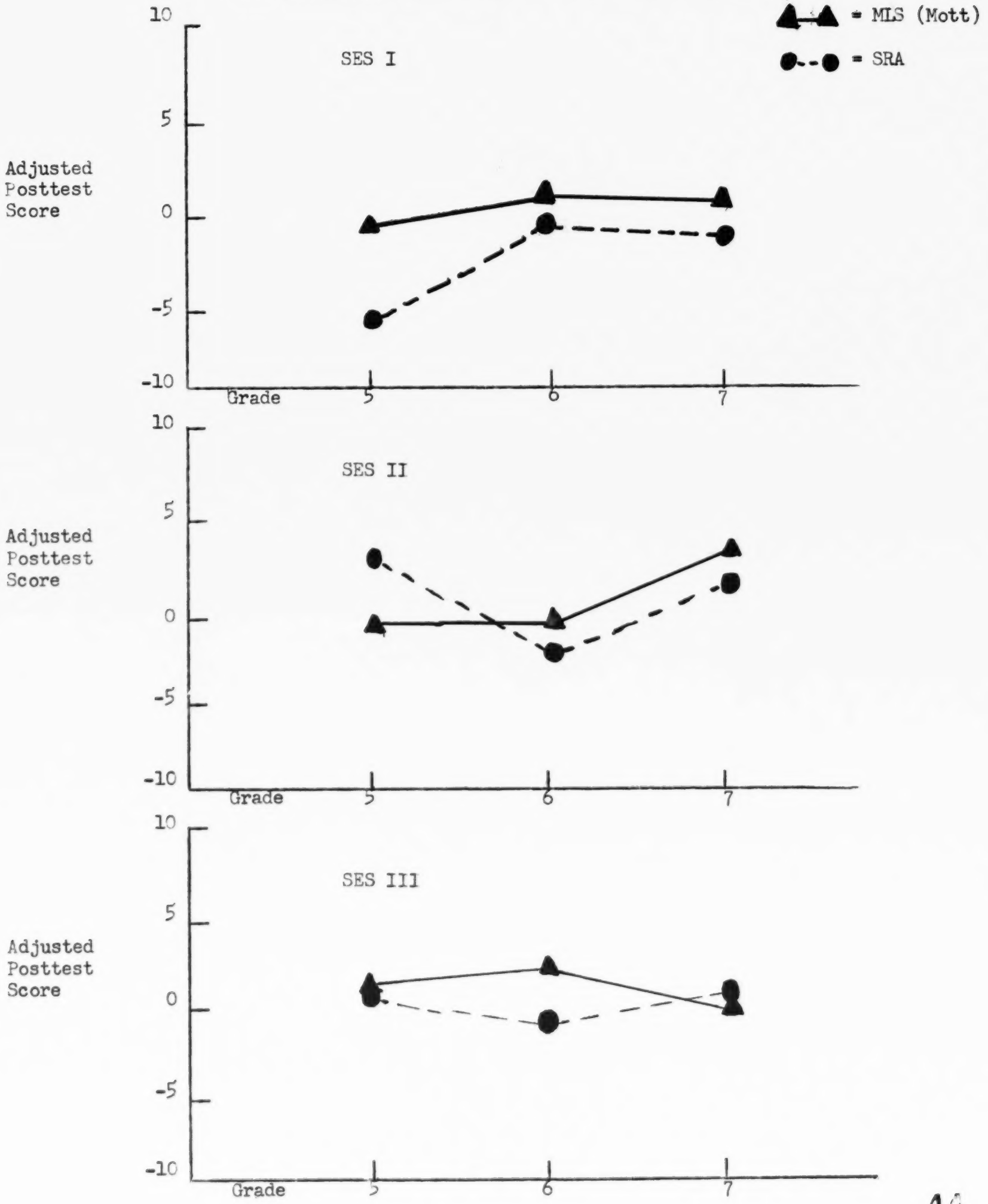

Figure 2.

Program Posttest Means Adjusted for Pretests
for SES x Grade x Treatment: Root Word

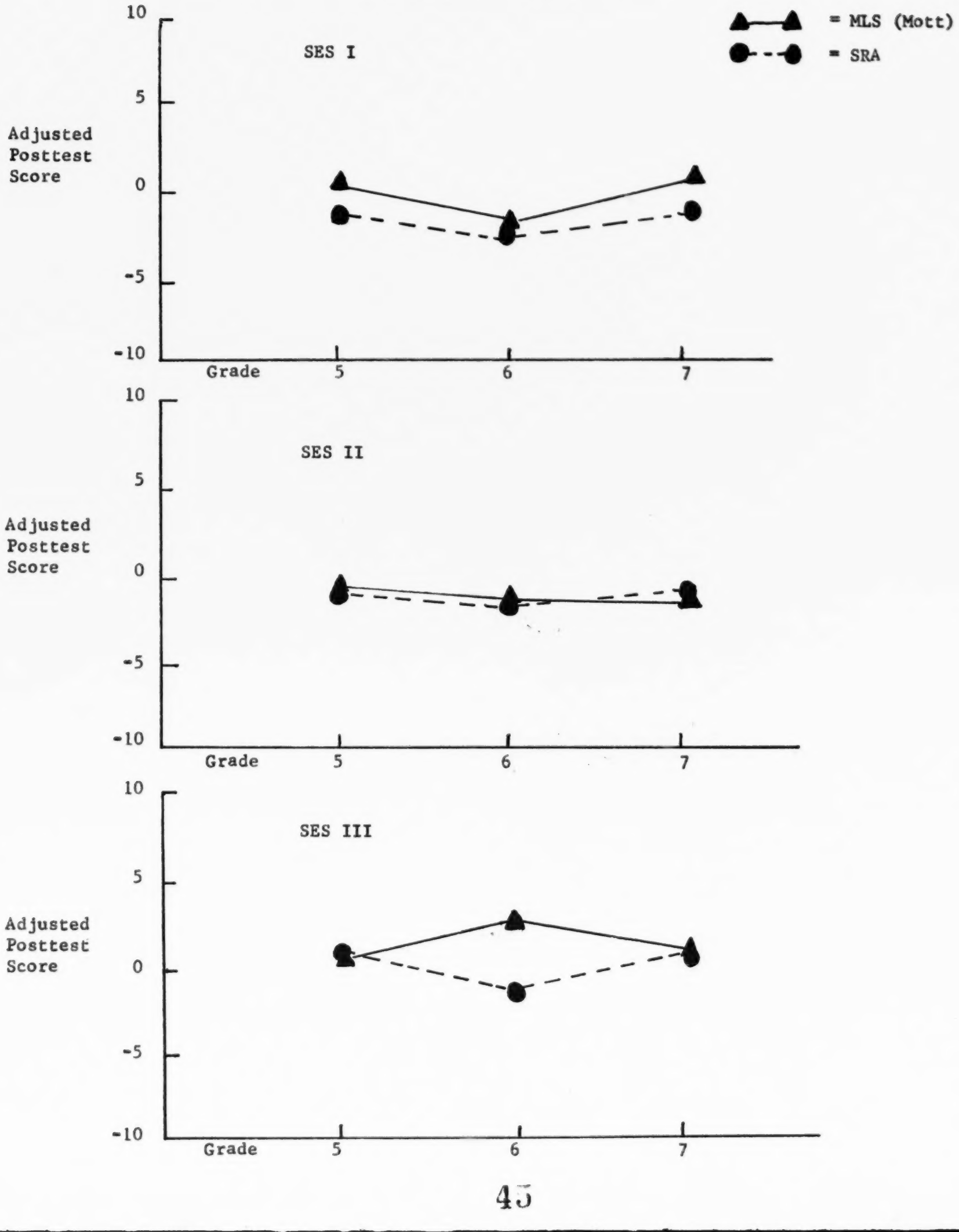

Figure 3.

Program Posttest Means Adjusted for Pretests
for SES x Grade x Treatment: Sound-to-Letter

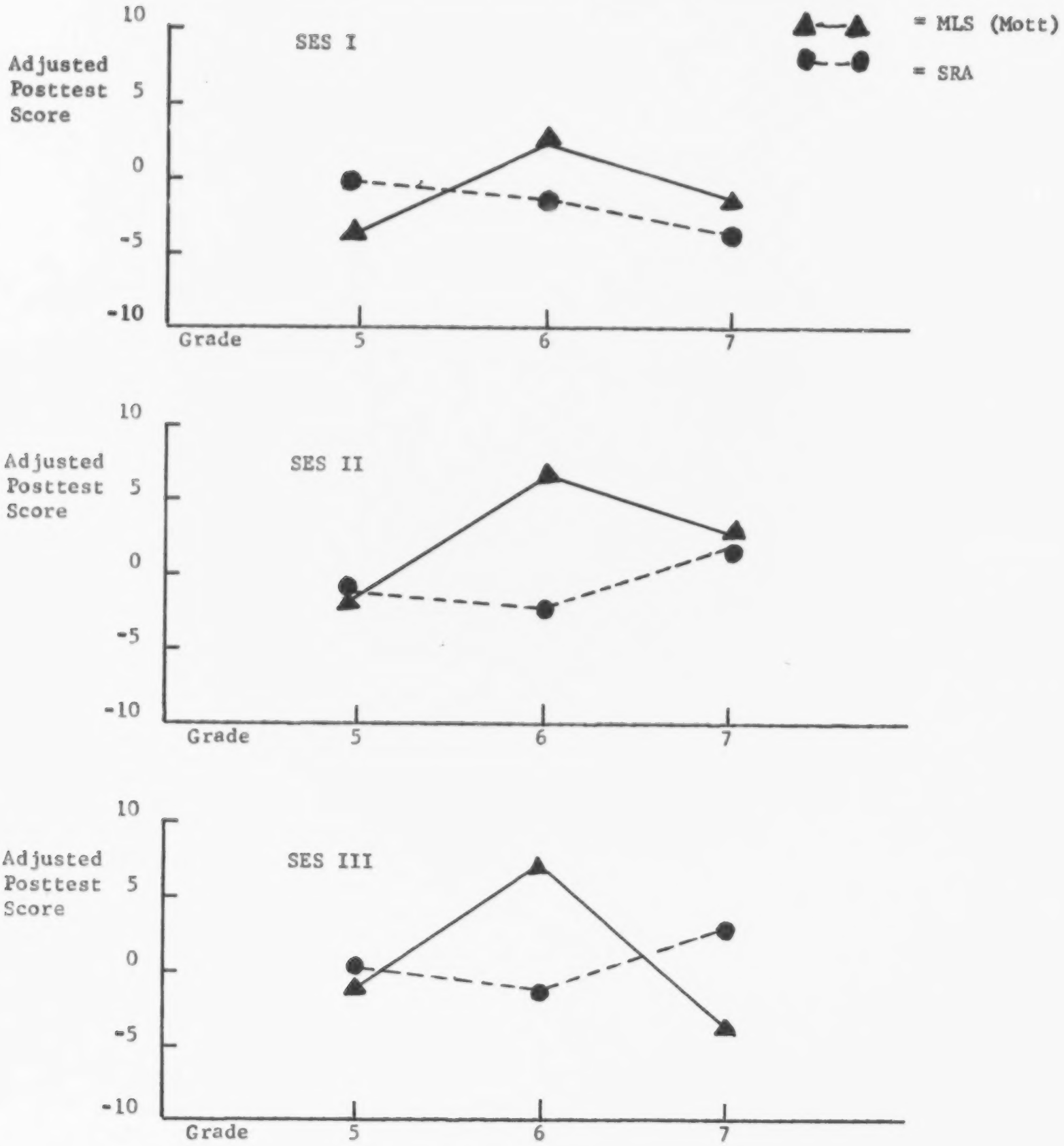

Figure 4.

Program Posttest Means Adjusted for Pretests
for SES x Grade x Treatment: Paragraph Comprehension

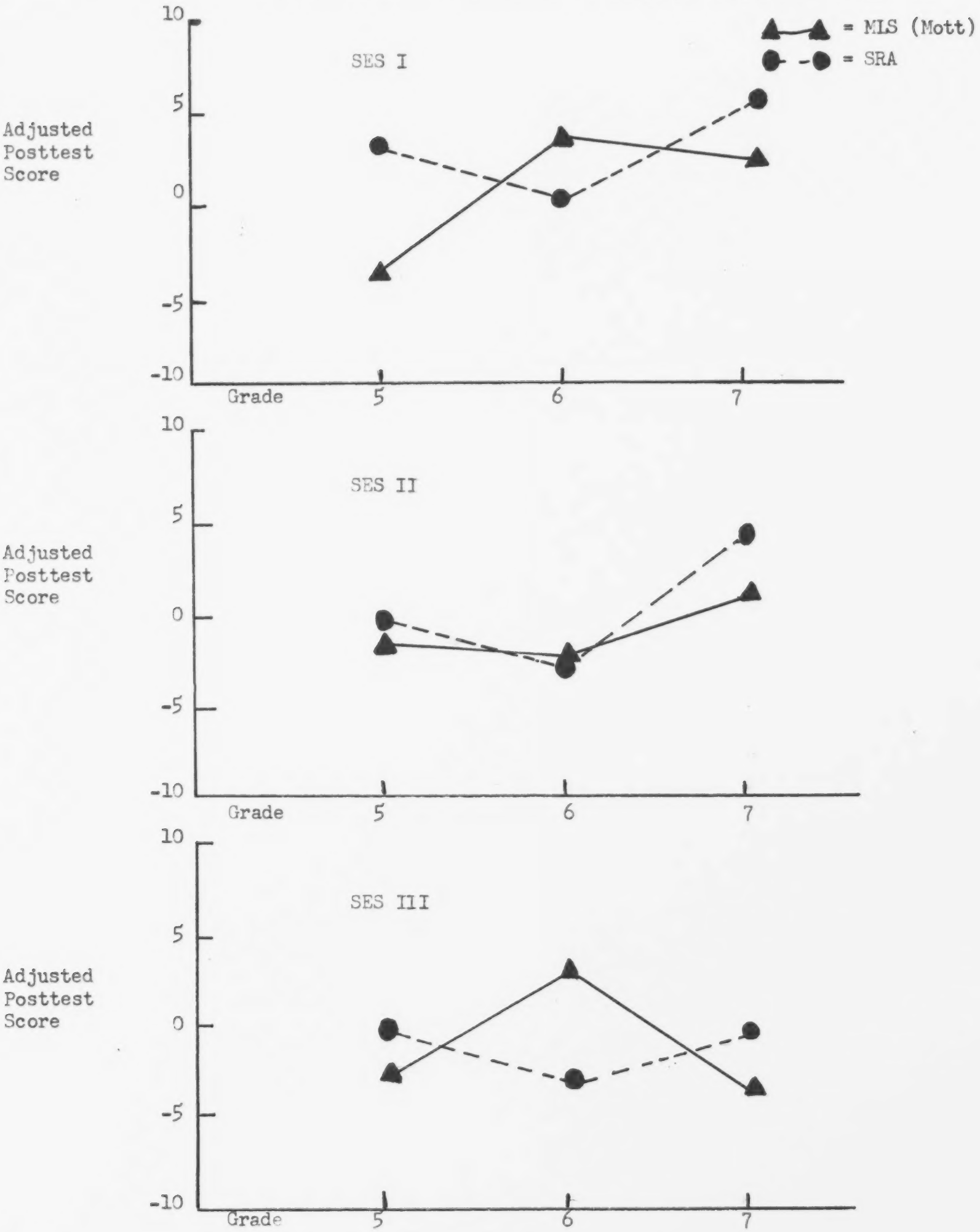

Figure 5.

Program Posttest Means Adjusted for Pretests
for SES x Grade x Treatment: Vocabulary

Figure 6.

Program Posttest Means Adjusted for Pretests
for SES x Grade x Treatment: Sentence Meaning

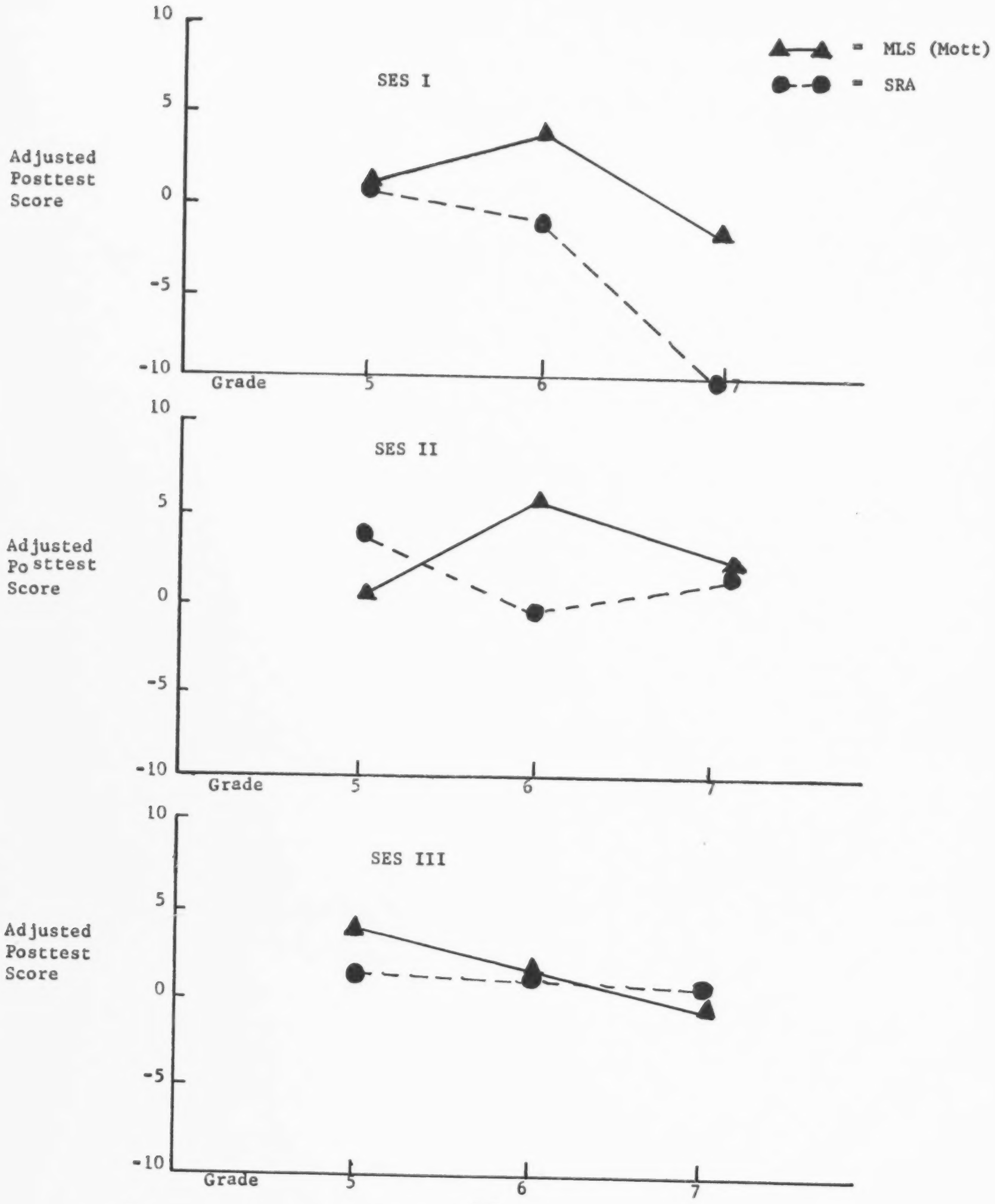

Figure 7.

Program Posttest Means Adjusted for Pretests
for SES x Grade x Treatment: Letter-to-Sound

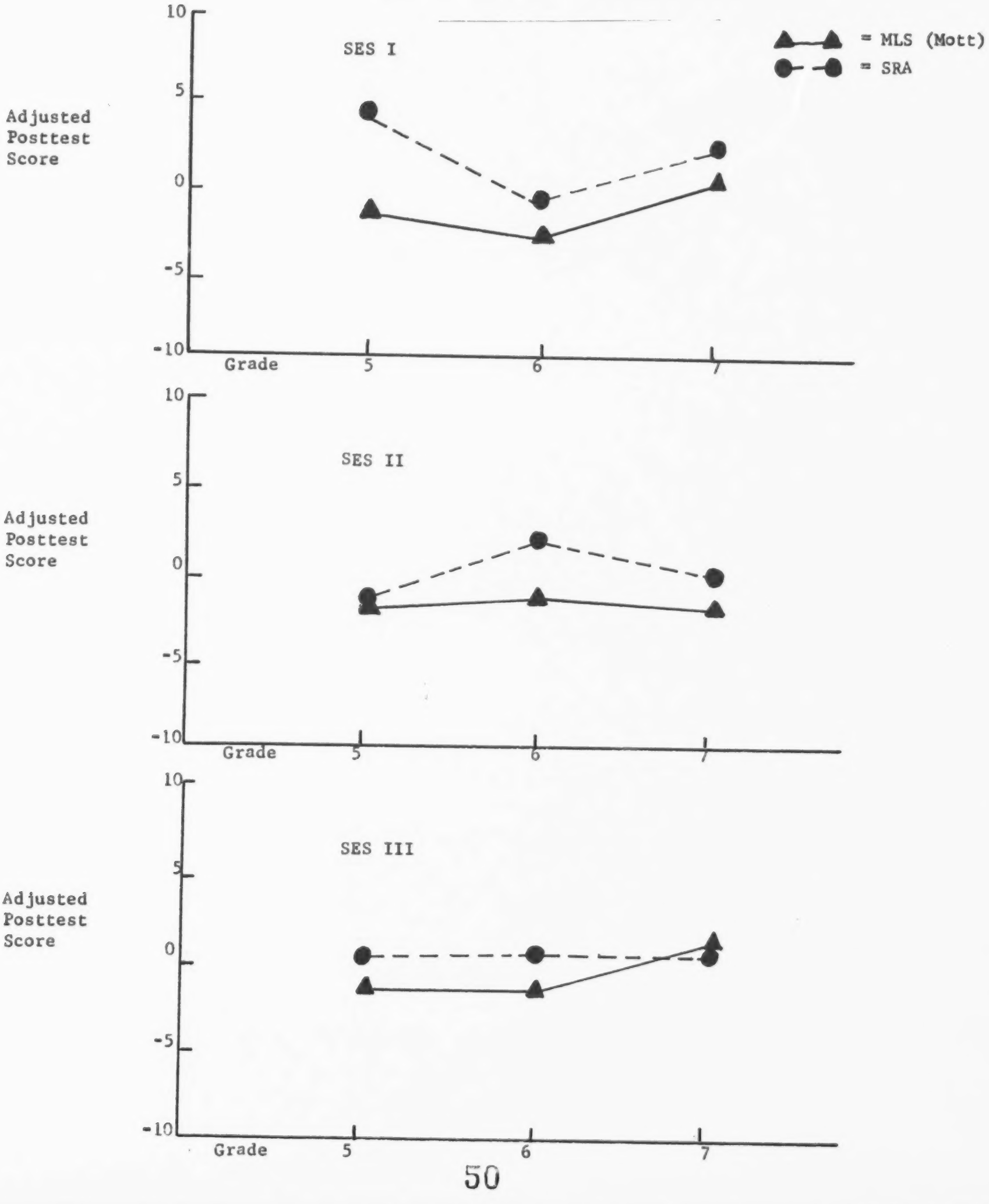

Figure 8.

Program Posttest Means Adjusted for Pretests
for SES x Grade x Treatment: Root Word

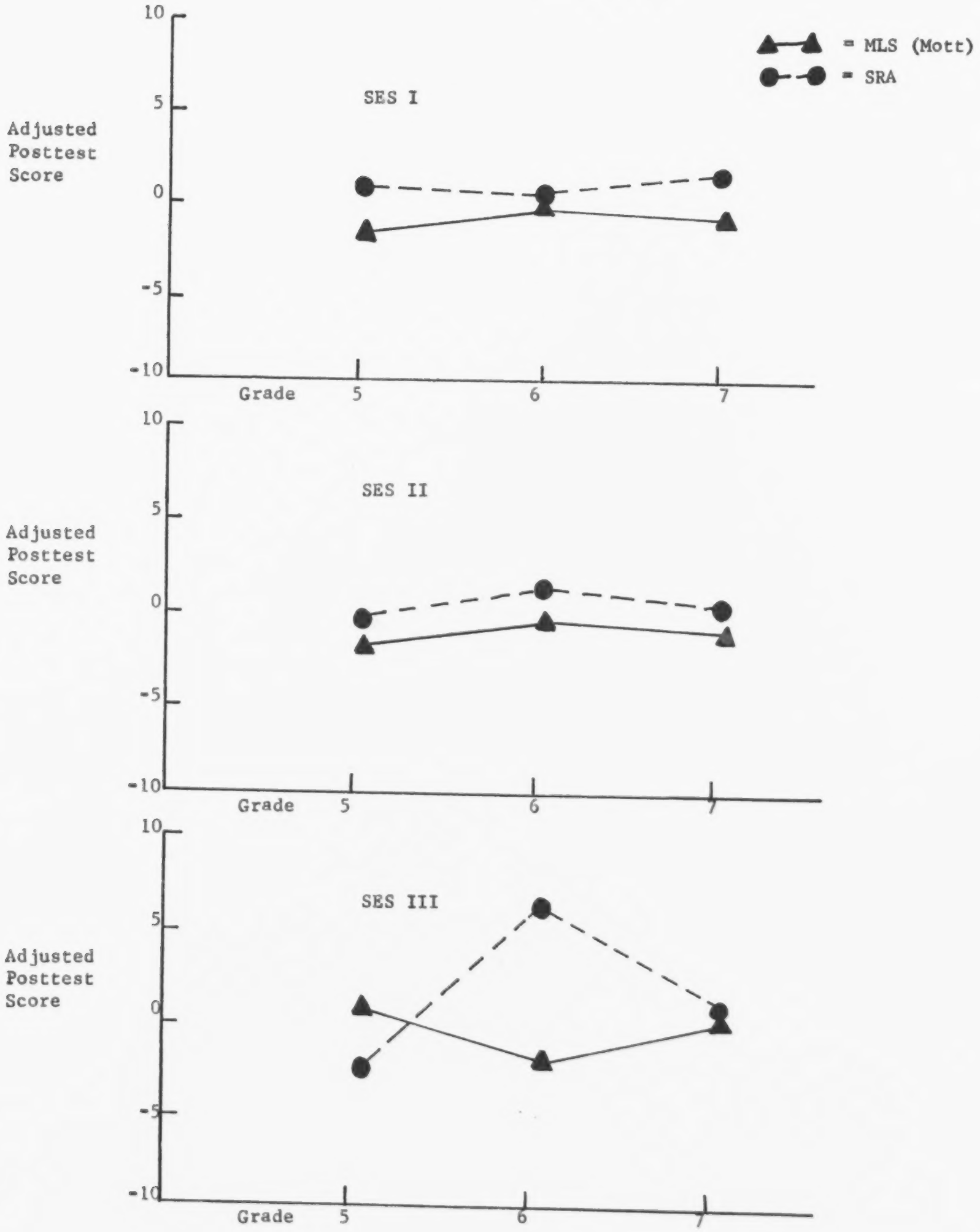

Figure 9.

Program Posttest Means Adjusted for Pretests
for SES x Grade x Treatment: Sound-to-Letter

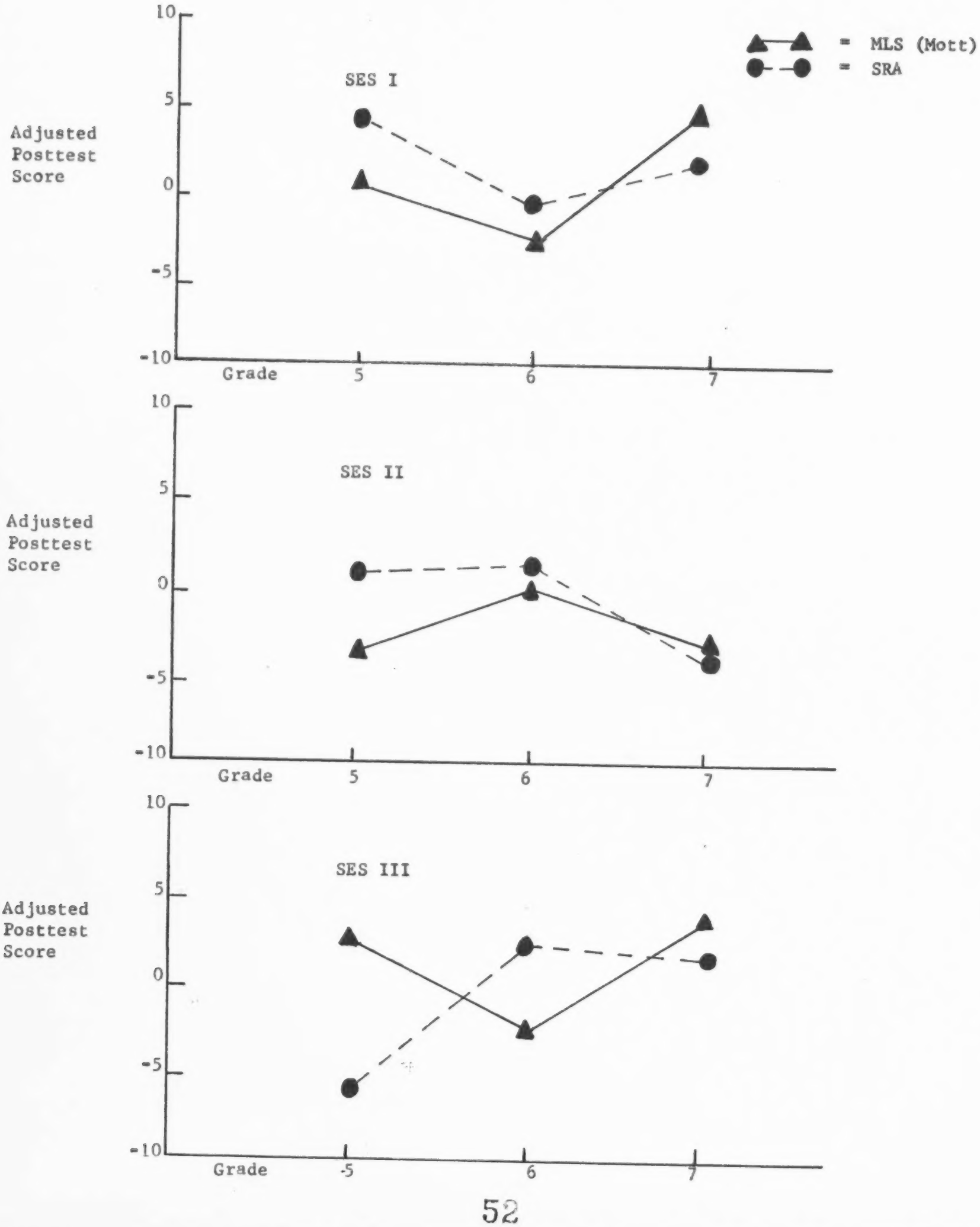

Figure 10.

Program Posttest Means Adjusted for Pretests
for SES x Grade x Treatment: Paragraph Comprehension

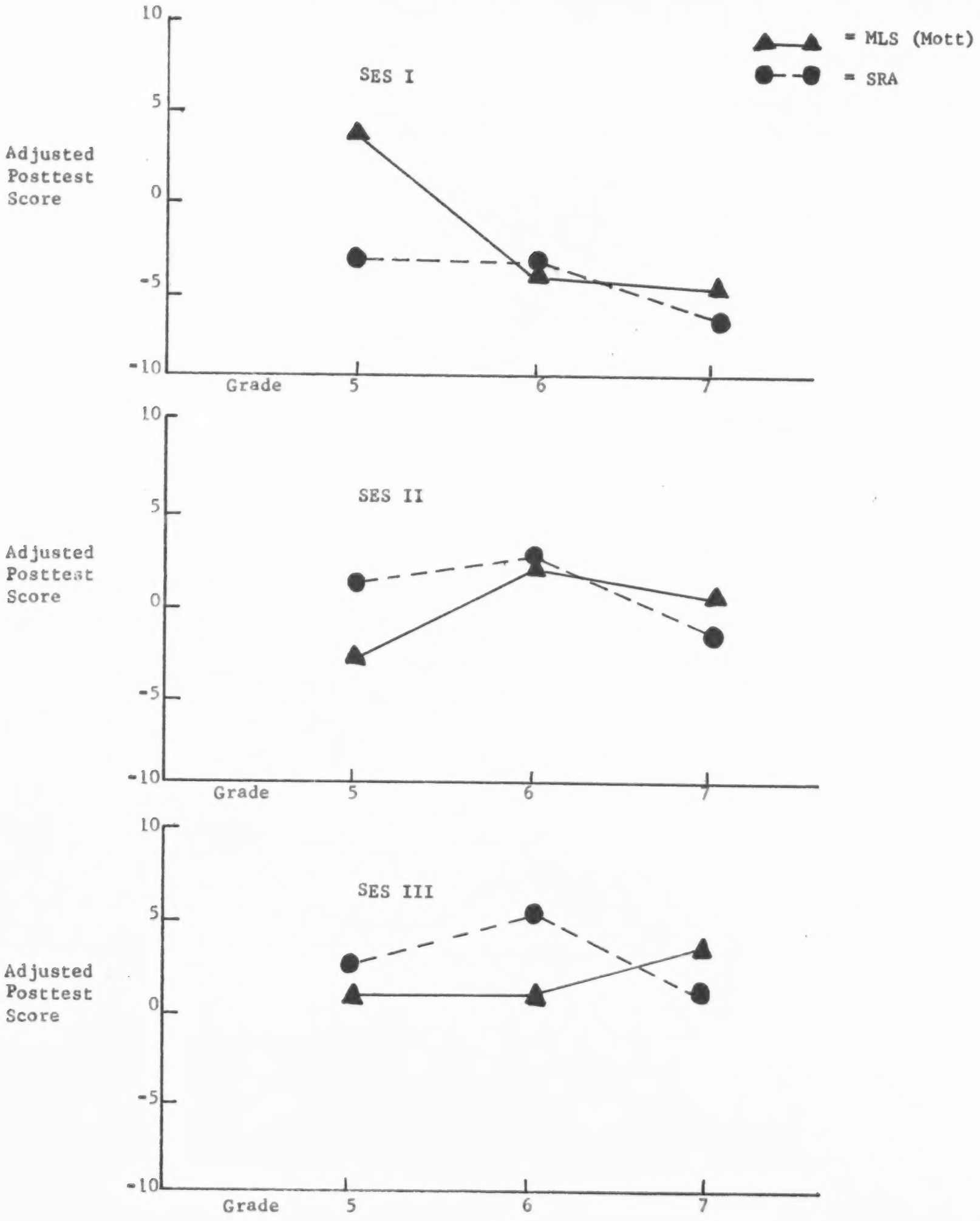

Figure 11.

Program Posttest Means Adjusted for Pretests
for SES x Grade x Treatment: Vocabulary

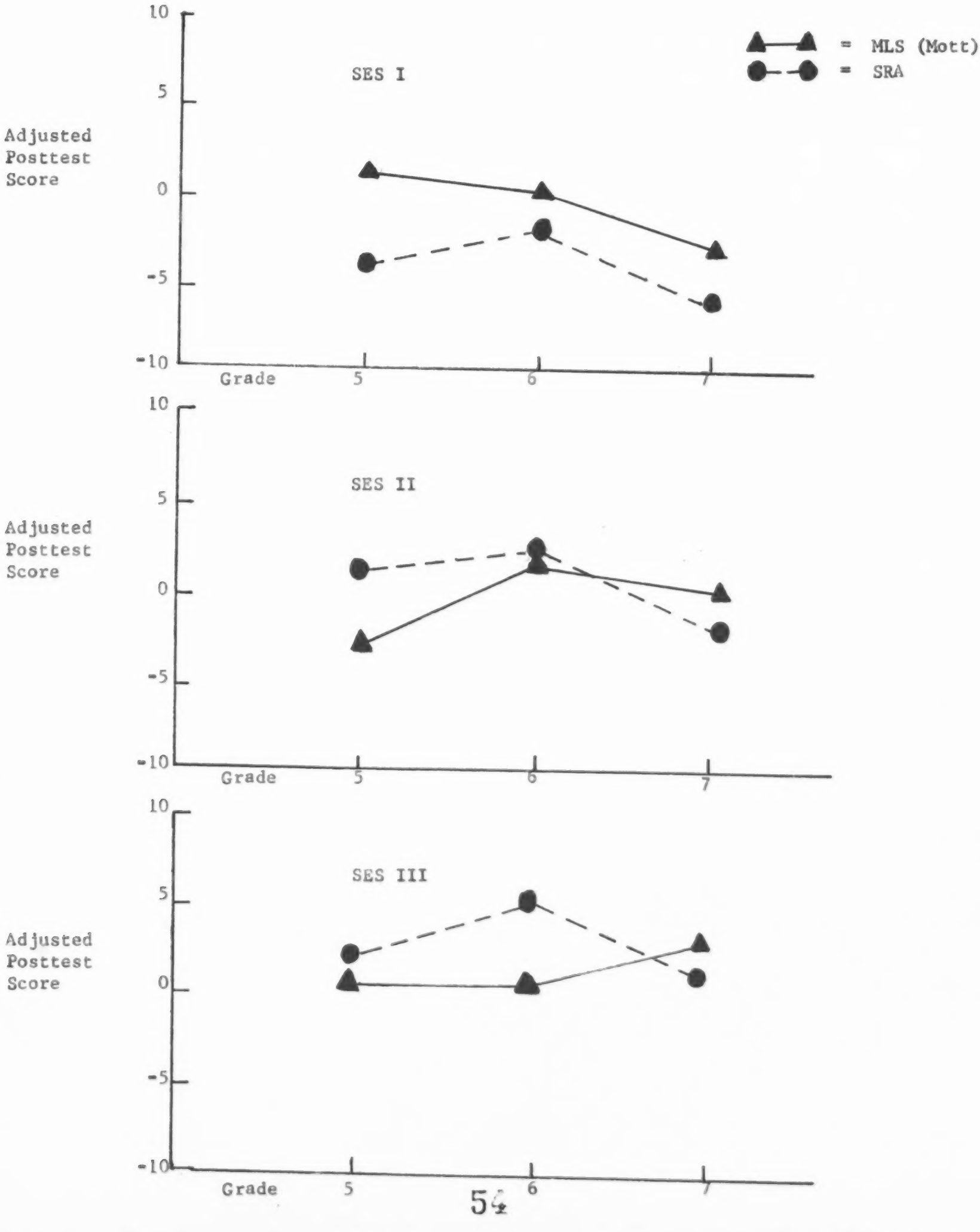

Figure 12.

Program Posttest Means Adjusted for Pretests
for SES x Grade x Treatment: Sentence Meaning

REFERENCES

Bormuth, J. On the Theory of Achievement Tests. Chicago: University of Chicago Press, 1970.

Cronbach, L. J. Course Improvement Through Evaluation. Teachers College Record, 64, 1963, 672-683.

Campbell, D. T. and Stanley, J. C. Experimental and Quasi-Experimental Designs for Research on Teaching. In N. L. Gage (ed.) Handbook of Research on Teaching. Chicago: Rand McNally, 1963.

Cronbach, L. J. and Snow, R. Individual Differences in Learning Abilities as a Function of Instructional Variables: Final Report. ERIC No. ED-029001, 1969.

Desberg, P. and Berdiansky, B. Word Attack Skills: Review of Literature. Southwest Regional Laboratory Report TR3, Inglewood, California, 1968.

Gagne, R. M. "Contributions of Learning to Human Development." Psychological Review, 75, 1968, 177-71.

Gagne, R. M. "Instructional Variables and Learning Outcomes." Center for the Study of Evaluation. Report No. 16, UCLA, 1969.

Jensen, A. "How Much Can We Boost IQ and Scholastic Achievement?" Harvard Educational Review, 39, 1969, 1-124.

Joreskog, K. G. "A General Method for Analysis of Covariance Structures with Applications: I and II." E. T. S. Bulletins RB-69-46 and 47, 1969.

Kempthorne, O. The Design and Analysis of Experiments. New York: Wiley and Sons, 1952.

Levin, H. and Gibson, E. J. (Eds.) Communicating by Language: The Reading Process. Proceedings of a conference sponsored by NICH, New Orleans, 1968.

Scriven, M. "The Methodology of Evaluation." In Perspectives of Curriculum Evaluation. Tyler, T., Gagne, R., and Scriven, M. Chicago: Rand-McNally, 1967.

Stake, R. Toward a Technology for the Evaluation of Educational Programs. In Tyler, Gagne, Scriven (eds.) Perspectives of Curriculum Evaluation. Chicago: Rand-McNally, 1967.

Tyler, R. W. Changing Concepts of Educational Evaluation. In Tyler, Gagne, Scriven (eds.) Perspectives of Curriculum Evaluation. Chicago: Rand-McNally, 1967.

Venezky, R. L., Calfee, R. C. and Chapman, R. S. Pronunciation of Synthetic words with Predictable and Unpredictable Letter Sound Correspondences. Technical Report No. 71, Wisconsin R. and D. Center, Madison, 1969.

Wiley, D. and Bock, R. D. Quasi-Experimentation in Educational Settings: Comment. School Review, 1967.

END